# Multiple testing on Experimental Economics Data

**Dongyin Hu**
Department of ECE
A59006035

## 1 Introduction

Benefited from the improvements in storage, communication and new technologies, the size of the data people have access to nowadays is ever growing. As reported in [KL21], the medical image research field has observed an growth of the geometric mean dataset size in an exponential trend. The researchers are more and more likely to deal with large dataset, and as well conducting many hypothesis tests to discover insights of the data. However, suppose each test is conducted at level $\alpha$ and the $p$-values are independent, then the family-wise error rate (FWER), the probability of one or more false rejections, equals $1-(1-\alpha)^N$ and grows quickly as the number of null hypotheses increases. Even with a small set of null hypotheses to test, for instance, $N=5$, the probability of making at least one false rejection is 22.64% when $\alpha=0.05$. At a larger scale, we are $99.4\%$ likely to make a false rejection when the size of total testing is $N=100$ and same $\alpha$. This is the multiple testing problem.

Many works have been proposed for controlling the FWER. From the well-known Bonferroni correction dated back in 1935 [Bon35], to a very recent work in 2019 [LSX19]. It is desired to keep the testing powerful as well as control the probability of making false rejections. In this project, the FWER control procedure proposed in [LSX19] and the theoretical framework proposed by [RW10] are applied on an economics data collected in 2007 as detailed in [KL07]. Some conclusions made in [KL07] are revisited and corrected with the procedure, as done in [LSX19]. Based on the updated version of the same data, where some household information is added, several insights not mentioned in the previous works are discovered.

The objectives of this project are:

1. Understand the FWER proposed in [LSX19] and replicate the results with a Python implementation of the algorithm;
2. Apply the procedure to investigate the data and make new conclusions.

This report is organized as follows: section 2 offers the background information about the procedure of data collection; section 3 discusses the theoretical explanations of the FWER control procedure and offers pseudocodes of the implementation; and section 4 reproduces the results in [KL07] and [LSX19], as well as discovers some new insights; finally, section 5 summaries the report.

The codes for the project are uploaded at here.

## 2 Background

The work [KL07] investigates a charitable fundraising problem. To increase the revenue, fundraisers often adopt matching gifts. A matching gift is a gift that is committed by a donor(s) conditioned on the contributions of others at a given rate, up to a maximum amount this leadership donor is prepared to give. A matching rate of $X:1$ means that for each dollar contributed by others, the leadership donor donates $X$ dollars. It is empirically believed that a higher matching rate have noticeable power to influence future contributions.

To examine this rule of thumb, researchers sent direct letter solicitations to 50,083 prior donors who have contributed to the Organization at least once since 1991. The letter discussed Supreme Court nominations. A number of 16,687 subjects, or 33% of the sample are randomly assigned as the control group. All donors received a four-page letter identical in all respects except the treatment letters included an additional information on the matching offer. The match offer is randomized with equal probability in three aspects:

1. The matching ratio ($1:1$, $2:1$, or $3:1$),
2. The maximum amount of the matching offer ($25,000, $50,000, $100,000 or unstated),
3. the suggested donation amount (equals, 1.25 times or 1.5 times the highest previous donation).

The response variables are 1) whether the donor contributed within a month after receiving the letter, 2) the dollars given without the matching amount (raw contribution), 3) the dollars given with the matching amount ($X+1$ times the raw contribution), and 4) the change in the amount given. The data are analyzed using Probit model (on variable 1) and OLS linear regression model with respect to the treatments (independent variables) on different subgroups in order to examine the heterogeneous treatment effects. Some of the conclusions from [KL07] are:

1. Using matching grants increases both the revenue per solicitation and the probability that an individual donates;
2. Larger match ratios (i.e., $3$ :1 and $2$ :1) relative to smaller match ratios ($1$ :1) have no additional impact;
3. The matching grant works for donors in red states but not blue states;

Later, one of the authors of [KL07] revisited the fundraising data with the FWER control procedure proposed in a new work [LSX19]. The core idea of the proposed procedure is to correct the $p$-values with the bootstrap estimation of its empirical cumulative density function (ECDF). In comparison with the Bonforroni procedure, this procedure considers dependence structure among $p$-values and thus is more powerful.

## 3 Theoretical Calculations

In this section we discuss the theoretical behind the control procedure. The notations used in the following subsections follow the spirit in [RW10] and [LSX19], yet may differ as this report aims to provide only the pertinent results. Some notations are simplified.

### 3.1 Notations

Assume the observed data $X^{(n)}$ is generated from some unknown distribution $P$, where $n$ is the sample size. A model assumes that $P$ belongs to a certain family of probability distributions $\Omega$. We would like to consider the problem of simultaneously testing a set of null hypotheses $H_i$ against alternative hypotheses $H_i'$, for $i = 1, \cdots, m$. The test statistics for $H_i$ is denoted as $\hat{T}_{n,i}$, where $i$ is the hypothesis index and $n$ denotes that is is calculated from $n$ samples. A large $\hat{T}_{n,i}$ indicates evidence against the null hypotheses $H_i$. Then, to control the family-wise error rate (FWER) is to control the following probability:

$$\text{FWER}_P = \mathbb{P}\left\{\text{Reject at least 1 hypothesis } H_i : i \in I(P)\right\} \leq \alpha,$$

where $I(P)$ is the index set of true null hypotheses if $P$ is the underlying distribution.

Let $T_i(P)$ be the test statistic under the real unknown distribution. Considering that testing whether a test statistic is large enough is equivalent to testing if the observation falls in the condence interval, we consider the difference between the estimator and the test statistic under null hypothesis $\hat{T}_{n,i} - T_i(P)$. This difference is a random variable since $\hat{T}_{n,i}$ is a random variable, and we use $J_{n,i}(x; P)$ to denote the cumulative density function (CDF) of $\hat{T}_{n,i} - T_i(P)$ under $P$, i.e., $J_{n,i}(x; P) = \mathbb{P}(\hat{T}_{n,i} - T_i(P) \leq x)$. Similarly, $H_{n,i}(x; P)$ denotes the CDF of the absolute difference $|\hat{T}_{n,i} - T_i(P)|$ under

$P$. With these notations, we can write the confidence interval as:

$$\left\{T_i : \left|\hat{T}_{n,i} - T_i\right| \le c_{n,i}(\gamma; P)\right\},$$

where $c_{n,i}(\gamma; P)$ is the largest $\gamma$ quantile of $H_{n,i}(x; P)$.

In order to control the FWER, we require:

$$\begin{aligned}
1 - \text{FWER} &= \mathbb{P}\left\{\left|\hat{T}_{n,i} - T_i(P)\right| \le c_{n,i}(\gamma; P) \text{ for all but at most } 1 \; i \in I(P)\right\} \\
&= \mathbb{P}\left\{H_{n,i}\left(\left|\hat{T}_{n,i} - T_i(P)\right|; P\right) \le \gamma \text{ for all but at most } 1 \; i \in I(P)\right\} \\
&= \mathbb{P}\left\{\max_{i \in I(P)} H_{n,i}\left(\left|\hat{T}_{n,i} - T_i(P)\right|; P\right) \le \gamma\right\} \\
&\ge 1 - \alpha,
\end{aligned}$$

where the solution of $\gamma$ is given by the $1 - \alpha$ quantile of the distribution of $\max_{i \in I(P)} H_{n,i}(|\hat{T}_{n,i} - T_i(P)|; P)$, which is then denoted by $L_{n,I(P)}(x; P)$.

### 3.2 Asymptotic Behavior of FWER, and Balance

Since $P$ is unknown, it is estimated by $\hat{Q}_n$, for example the ECDF when no parametric model for $P$ is provided. All above statistics can be estimated by plugging in $\hat{Q}_n$. Thus, the distributions, like $H_{n,i}(x; P)$ and $L_{n,I(P)}(x; P)$, can be estimated using the bootstrap method.

To be specific, for each $i$, we draw a sample of size $n$ from $\hat{Q}_n$ for $B$ times. Each time we can calculate an estimate of $T_i(P)$, denoted as $\hat{T}_{n,i}(b)$, for $b = 1, \cdots, B$. We can then estimate the previous two distribution by:

$$\begin{aligned}
H_{n,i}(x; \hat{Q}_n) &= \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\left\{\left|\hat{T}_{n,i} - \hat{T}_{n,i}(b)\right| \le x\right\} \\
L_{n,I(P)}(x; \hat{Q}_n) &= \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\left\{\max_{i \in I(P)}\left|\hat{T}_{n,i} - \hat{T}_{n,i}(b)\right| \le x\right\},
\end{aligned}$$

where $\mathbf{1}$ is the indicator function. The total complexity is $O(mB)$ for estimating both distribution, which is linear with respect to the total number of hypotheses and the bootstrap times, respectively. This thanks to the reuse of random samples when estimating $L_{n,I(P)}(x; P)$.

The confidence interval for each individual null hypotheses $H_i$ constructed by using $\hat{Q}_n$ is:

$$\hat{c}_{n,i} = \hat{T}_{n,i} \pm H_{n,i}^{-1}\left(L_{n,I(P)}^{-1}\left(1 - \alpha; \hat{Q}_n\right); \hat{Q}_n\right), \tag{1}$$

which offers simultaneously coverage for all the true test statistic $T_i(P)$, except for at most 1 of them, with an asymptotic probability of $1 - \alpha$, when certain assumptions are met. This is equivalent to write as:

$$\lim_{n \to \infty} \text{FWER} \le \alpha \tag{2}$$

The assumptions are about the underlying distribution $P$ should be nondegenerate, have connected support. In addition, the bootstrap procedure needs to be consistent. For a detailed description and proof please refer to [RW10] and [LSX19].

In addition, the intervals are balanced. Since in equation 1 the confidence interval is dependent on $i$, each individual test statistic is covered with the same probability $\gamma$. This results:

$$\lim_{n \to \infty} \mathbb{P}\left\{T_i(P) \in \hat{c}_{n,i}\right\} = \gamma, \; \forall i \in 1, \cdots, m \tag{3}$$

The requirement of balance is due to multiple outcomes are considered. Some outcomes may take much larger values then other outcomes. If no balance requirement is imposed, some true null hypotheses with a larger outcome scale may be likely to be rejected than other true null hypotheses. Alternatively, balance can sometimes be achieved by studentization. A detailed discussion is available in [RW10].

**Algorithm 1:** The step-down FWER control procedure.

```
   Input: Samples $X^{(n)}$, FWER bound $\alpha$, Number of Null Hypotheses $m$
   Output: The index set of accepted hypotheses $S$
1  $S \leftarrow \{1, \cdots, m\}$;
2  while $S$ is not empty do
3      $t \leftarrow []$;
4      for $s$ in $S$ do
5          Use bootstrap to get $H_{n,s}(\cdot; P)$;
6          $t.append(H_{n,s}(|\hat{T}_{n,s}|))$ ;
7      end
8      Reuse bootstrap results to get $L_{n,s}(\cdot; P)$;
9      $c \leftarrow L_{n,S}^{-1}(1 - \alpha; \hat{Q}_n)$;
10     if $t.max() \leq c$ (no hypotheses is rejected) then
11         break;
12     end
13     $S \leftarrow \{s \in S : t[s] \leq c\}$;
14 end
15 return $S$
```

### 3.3 The Step-down Algorithm

Often, single-step methods can be improved in terms of power via stepwise methods, while nevertheless maintaining control of the desired error rate [RW10].

Let $S \subseteq \{1, \cdots, m\}$, where $m$ is the total number of testing. Denote $L_{n,S}(x; P)$ as the CDF of $\max_{i \in I(P)} H_{n,i}(|\hat{T}_{n,i} - T_i(P)|; P)$ under $P$, i.e.,

$$L_{n,S}(x; P) = \mathbb{P}\left(\max_{i \in S} H_{n,i}\left(|\hat{T}_{n,i} - T_i(P)|; P\right) \leq x\right), \tag{4}$$

where the difference is the index set to search the maximum over. Further, we assume that under null hypothesis $T_i(P) = 0$. Then, the algorithm to control the FWER as well as improve the power of the test is noted as algorithm 1 following [RW10] and [LSX19]. The dependence between the test statistics is implicitly taked into account by the bootstrap procedure, as [LSX19] mentioned.

There are several variants from algorithm 1.

1. Instead of return the set of accepted null hypotheses, the algorithm can return the adjusted $p$-values for each hypotheses by using the smallest value of $\alpha$ for which $H_s$ is rejected (Remark 3.6 of [LSX19]). As for implementation, we sort the $p$-values in ascending order, and correct by calculate the smallest value of $\alpha$ for each $p$-value.
2. By replacing the critical value $c$ in line 9 of algorithm 1, some other control procedure can be derived. (Remark 3.2 of [LSX19]). For example, to achieve Bonferroni procedure, we replace it with $1 - \alpha/|S|$.

## 4 Full data analysis

We first revisit the results offered in [KL07] and [LSX19]. Then we apply the similar procedure to discover some other results not presented in neither [KL07] nor [LSX19]

Even though the main focus of the project is applying the $p$-value adjustment procedure proposed in [LSX19], the generalized linear models (Probit) and OLS linear regression models are applied as a complementary analysis to compare with.

### 4.1 Reproduction of Results

This part **reproduces** the results in [KL07] and [LSX19]. This is necessary as the FWER control procedure proposed by [LSX19] is re-implemented in Python, while the original codes are in either

Matlab or Stata. In addition, the analysis codes for [KL07] are in Stata as well, and they are translated into R codes for the following results shown in the report as to understand and practice coding in R and confirm the provided results. In addition, some notations are reused in later parts.

### 4.1.1 Regression Analysis

Two different models are applied in order to analyze different outcomes as in [KL07]. Results in here are reproduced using R codes rather than original Stata codes.

For analyzing the response rate, it is equivalent to model this by a binary random variable following the Bernoulli distribution where being 1 represents respond. The Probit model is used, which is a generalized linear model takes the form:

$$\mathbb{P}\{Y=1|X\} = \Phi(X^\top \beta),$$

where $\Phi$ is CDF of the standard normal distribution, $\beta$ is the parameter. Two specifications over all subjects are estimated:

$$\begin{aligned} Y_i &= \Phi(\beta_0 + \beta_1 T_i + \epsilon_i) \\ Y_i &= \Phi(\beta_0 + \beta_1 T_i S_i + \beta_2 T_i P_i + \beta_3 T_i X_i + \epsilon_i), \end{aligned}$$

where $i$ is the subject index, $Y_i$ is the binary random variable representing responding or not, $T_i$ is the indicator variable of whether the subject is in the treatment group or not, $S_i$ is a vector of three indicator variables for three of the four match sizes (the omitted category is unstated). $P_i$ is a vector of two indicator variables for two of the three price ratios (the omitted category is $1:1$). $X_i$ is a vector of two indicator variables for two of the three example amounts (the omitted category is the low example amount). The omitted settings are estimated in the intercept term.

For analyzing the dollars given not including match, a OLS linear model is used. Similarly, two specifications over all subjects are estimated:

$$\begin{aligned} A_i &= \beta_0 + \beta_1 T_i + \epsilon_i \\ A_i &= \beta_0 + \beta_1 T_i S_i + \beta_2 T_i P_i + \beta_3 T_i X_i + \epsilon_i, \end{aligned}$$

where $A_i$ is the continuous random variable representing the dollars donated not including matching. Again, the omitted settings are estimated in the intercept term.

Please refer the R markdown file at here for the results. They are omitted in the report as it is not the main focus of the report and it is not related with new insights. Here we summaries the conclusions made in [KL07] using the language of [LSX19]:

1. Multiple outcomes case: Using matching grants increases both the revenue per solicitation and the probability that an individual donates;
2. Multiple treatments case: Larger match ratios (i.e., 3 :1 and 2 :1) relative to smaller match ratios (1 :1) have no additional impact;
3. Multiple subgroups case: The matching grant works for donors in red states but not blue states;

### 4.1.2 Multiple Hypothesis Testing

To better represents which treatment, subgroup, and outcome is being tested, it is better to replace the integer index with a four-tuple as in

$$S = \{(d, d', z, k),\, d, d' \in D, z \in Z, k \in K\},$$

where $D$ is the set of treatment status, $Z$ is the set of all subgroups, and $K$ is the set of outcome indices. The null hypotheses is then:

$$H_s : \mathbb{E}_P[Y_{i,k}(d) - Y_{i,k}(d')|Z_i = z] = 0,\ s = (d, d', z, k) \in S \tag{5}$$

where $Y_{i,k}(d)$ is the $k$-th outcome of subject $i$ when the treatment $d$ is applied. The test statistics used is:

$$\hat{T}_{n,s} = \sqrt{n}\left| \frac{1}{|I_{d,z}|} \sum_{i \in I_{d,z}} \left(Y_{i,k} - \hat{\mu}_{k|d,z}\right) - \frac{1}{|I_{d',z}|} \sum_{i \in I_{d',z}} \left(Y_{i,k} - \hat{\mu}_{k|d',z}\right) \right|,$$

where $I_{d,z} = \{1 \leq i \leq n : D_i = d, Z_i = z\}$ is the subject index set for subjects under treatment $d$ and subgroup $z$, $|I_{d,z}|$ is the cardinality of $I_{d,z}$, and $\hat{\mu}_{k|d,z} = \frac{1}{|I_{d,z}|} \sum_{i \in I_{d,z}} Y_{i,k}$ is the mean of the $k$-th outcome for the treatment $d$ and subgroup $z$.

This statistic is an estimator of the statistic in 5, which measure whether the $k$-th outcome is different between the group of subjects with treatment $d$ and subgroup $z$ and the group of subjects with treatment $d'$ and subgroup $z$. Note that under the null hypotheses, this statistic should be $0$, so we can omit $T_{n,s}(P)$ in the equations mentioned in Section 3.

The results obtained by following the same testings done in [LSX19] are listed in 1 and **??**. They are almost identical to the original results, except for 1) the way quantiles are calculated (to be specific, the 0-th quantile), which influences the adjusted $p$-value by $1/B$ (negligible when $B$ is large); 2) the results with respect to blue states as in Panel B. Even the absolute difference which does not involves further computation does not match with the original results. This might be a problem with the dataset itself, though if only statistical significance are checked, the conclusions are still the same.

**Multiple outcomes case**. First we revisit whether matching grants encourages receivers to donate. This is reproducing the results in [LSX19]. The set of null hypotheses in this case is indexed by the set:

$$S_1 = \{(d, d', z, k),\ d, d' \in D_1, z \in Z_1, k \in K_1\}\,, \tag{6}$$

where $D_1 = \{\texttt{Treatment}, \texttt{Control}\}$ means we compare treatment group and control group, $Z_1 = \{\texttt{All}\}$ since we do consider subgroups in this case, and $K_1 = \{$`Respond Rate`, `Dollars Given Not Including Match`, `Dollars Given Including Match`, `Amount Change`$\}$ since we are examining all four outcomes. The total number of hypotheses is $|S_1| = 4$.

Panel A of table 1 lists the $p$-values of the tests. Without correction, the tests strongly suggest that the treatment influences the response rate and dollars given including match, mildly suggest that dollars given not including match is affected, and the amount change is not influenced. The correction following equation 1 holds the same conclusion.

Note that Bonferroni and Holm's corrections show that the influence of matching grants on the dollars given not including match is not significant. This proves that the FWER control procedure in [LSX19] is more powerful than these two corrections.

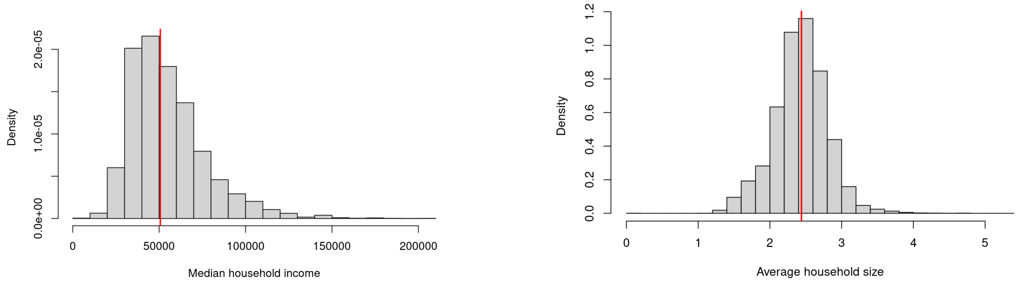
**Multiple subgroups case**. This parts we examines whether the treatment influences the response rate for people residing in different states and counties. Similarly, the indices set $S_2$ is characterized by $D_2 = \{\texttt{Treatment}, \texttt{Control}\}$, $Z_2 = \{$`Red county in a red state`, `Blue county in a red state`, `Red county in a blue state`, `Blue county in a blue state`$\}$ representing the four combinations of red/blue states and red/blue counties, and $K_2 = \{$`Respond Rate`$\}$. The total number of hypotheses is $|S_2| = 4$.

The raw $p$-values in Panel B of table 1 suggest that matching grants are working effectively for red counties in red states strongly, for blue counties in red states mildly, and not effectively in blue states. Yet after correction, all three corrections indicate that for blue counties in red states the matching grants has no significant affect.

**Multiple treatments case**. This parts we examines whether the different matching ratios affects the dollars given without match. Similarly, the indices set $S_3$ is characterized by $D_3 = \{$`Control,1:1,2:1,3:1`$\}$, $Z_3 = \{\texttt{All}\}$ since no subgroup is considered, and $K_3\{$`Dollars given not including match`$\}$. The total number of hypotheses is $|S_3| = 6$.

From results before correction in Panel C of table 1, $2 : 1$ is found to be significant in the dollars given without match, which implies this ratio might be optimal. Yet after correction, we found this significance disappear, and different matching ratios have no significantly different effect. This contradicts with the rule of thumb used by fundraisers.

**Multiple outcomes, subgroups, and treatments**. The four outcomes, four subgroups, and three treatments described before are combined together and tested jointly. There are a total of 48 null hypotheses of interest. As shown in table 2, 21 null hypotheses are rejected at the 10% level before correction, and the number reduced to 9 after correction at the same level.



(a) Distribution of median household income.

(b) Distribution of average household size.

Figure 1: The distribution of new variables to be analyzed. Red lines shows the median, which is used for dividing subjects into different groups.

### 4.2 New Insights

We further try to discover some new insights on an updated version of the same dataset, with more information available including the average household size and the median household income. We examine heterogeneous treatment effects between two outcomes (response rate and dollar given not including match) and treatments based on the following groups:

1. different genders;
2. different household income levels;
3. different household size levels.

For levels of household income, we use the median of median household incomes of all subjects as the threshold. Subjects with a median household incomes higher than the threshold are categorized as the relatively high median household income (high MHI) group, and subjects with that lower than the threshold are categorized as the relatively low median household income (low MHI) group. Similarly, for household size, we use the median of average household size of all subjects as the threshold and divide subjects into the group with relatively high or low average household size (high/low AHS). The similar grouping method is used in [KL07].

Figure 1 shows the distribution of the two new variables to study heterogeneous treatment effects.

The conclusions and procedure are **not** included in previous works.

#### 4.2.1 Regression Analysis

The specifications used in 4.1.1 is again adopted here, yet the subgroups are changed to the previous mentioned six subgroups.

Note that the when the independent variables are categorical, a linear regression model is equivalent as ANOVA [1]. According to the assumptions required by ANOVA [2], it is important for the sample sizes for each factor level to be equal, which is the case in this dataset as treatment is randomly assigned. Thus, the ANOVA results should be convincing.

The numbers in the table 3 are the marginal effect of each variable. Marginal effects tells us how a dependent variable (outcome) changes when a specific independent variable (explanatory variable) changes, which can be understood as a derivative. The Probit model shows the similar results as in male subjects, subjects from families with relatively low median household income are influenced by matching gifts on response rate. Further, a 2:1 ratio has statistically significant influence for families with relatively small household size on response rate. The OLS model shows that dollars given not including match is influenced by matching gifts for male subjects, subjects from families with relatively low median household income at 10% level.

[1] ANOVA As a Linear Model

[2] ANOVA Assumptions

### 4.2.2 Multiple Hypothesis Testing

This table shows the ordinary least squares LR results. We regress the amount of dollar given w.r.t. the treatments. Similar as the unadjusted results, the treatment influences the male and families with low MHI. But recall that the correction procedure actually do not reject these null hypotheses.

The null hypotheses format and same test statistic in 4.1.2.

We found that male subjects, subjects from families with relatively low median household income are influenced by matching gifts on response rate at 1% level, and families with relatively small household size is influenced at 5% level.

Without correction, the previous three groups are found influenced by matching gifts on dollars given not including match at 10% level. Yet all three corrections (using 1, Bonferroni or Holm's methods) do not provide evidence.

### 4.2.3 Explanations for Results in Multiple Hypothesis Testing

**Gender**. The total number of female subjects is 13,598 and that of male is 35,374. The response rates for female are $2.035\%$ (control) and $2.327\%$ (treatment), and for male the rates are $1.672\%$ (control) and $2.186\%$ (treatment).

Many studies have been investigating the gender generosity, yet the willingness to donate is usually regulated by the environment of the solicitation [3]. Considering the time when the experiment was conducted, it seems the nomination is about Samuel Alito (2006, one year before [KL07] published). This should be helpful to explain a higher response rate for women in the control group.

As for the explanation of larger increase in male, this might be related with man give from a strategic mindset as in [4] concluded. Using a simplified version of the utility model mentioned in [KL07], the utility an agent obtain from donation can be written as:

$$U_i = \delta_i h((X+1)b_i) + \gamma_i f(b_i) + c_i, \tag{7}$$

where $U_i$ is the utility obtained from donation, $f, h$ are non-decreasing concave functions that captures the relationship between utility and 1) the amount of donation, and 2) the total raised amount of donation as public good, respectively. $b_i$ is the amount of donation made by one, $X$ is the matching ratio $X : 1$, and $c_i$ is some other terms influencing the utility but omitted here. The parameter $\gamma_i$ might depend on the presence, and magnitude, of the promised matching grant monies. For male, this is potentially larger as it satisfies more about using "strategic mindset", and leading to a significant increase in the respond rate.

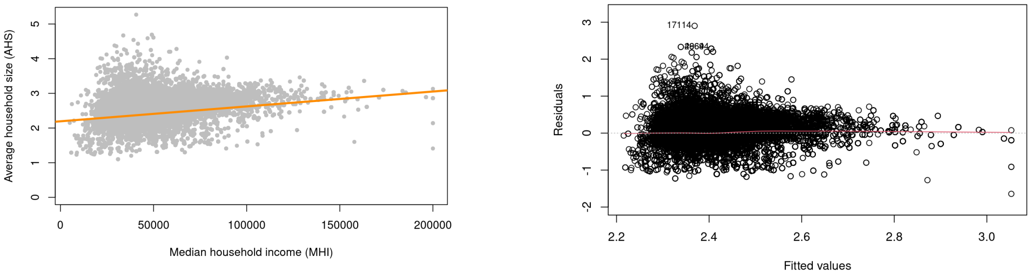
Please note this model is not correct since it assumes different matching ratios have different influence on the donation behaviours (regulated by utility), which is rejected by the testing in 4.1.2. A proper modification would be replace $X$ by a constant, so that the actual ratio does not influence the utility.

**Median household income**. This can be explained with 7 as well. Families with relatively median household income are usually more sensitive to the utility obtained per dollar. A matching gift increases $h((X+1)b_i)$ (note as well this might be incorrect since it assumes different ratios have different influence, but we can treat $X$ as a constant) and then increases the utility obtained per dollar. This drives families with relatively lower income more and thus we detect a significant increase in respond rate.

**Average household size**. This is somehow tricky to explain. At first, I assume the median household income and average household size are correlated. After model the average household size with median household income as independent variable, it seems when the median household income is relatively high, a linear relationship is clear while this relationship is less dominant when median household income drops. The visualization is available in 2. The evidence for contributing the significance of treatment on families with relatively low household size to its correlation with certain level of median household income. Maybe the reason comes from the environment of ask, which is possibly more pertinent to small-sized families.

[3] Gender Generosity: Who Gives More, And Does It Even Matter?

[4] Women and Giving. The impact of generation and gender on philanthropy.



(a) MHI ∼ AHS

(b) Residual vs. Fitted

Figure 2: Regression results for linear regression with median household income (MHI) as dependent variable and average household size as independent variable. When MHI is large, the linearity is clear.

## 5 Conclusions

The goals of 1) understanding the FWER procedure in [LSX19], 2) implementing the codes, and 3) discover new insights are met. This section summaries the whole report.

### 5.1 Findings

After applying the correction method following 1, the conclusions made in [KL07] are corrected as (this part is reproduction):

1. Matching grants influences response rate and dollars given including match at 1% level, and dollars given not including match at 10% level.
2. Matching grants influences subjects living in red counties in a red state at 10% level;
3. Different matching ratios are not significantly different.

The results above also show that the proposed FWER procedure in [LSX19] is more powerful than the traditional Bonferroni and Holm's methods.

In addition, some new insights from the updated version of data:

1. Male subjects, subjects from families with relatively low median household income are influenced by matching gifts on response rate at 1% level, and families with relatively small household size is influenced at 5% level;
2. Without correction, the previous three groups are found influenced by matching gifts on dollars given not including match at 10% level. Yet all three corrections (using 1, Bonferroni or Holm's methods) do not provide evidence.
3. The Probit model shows the similar results as in 1). Further, a 2:1 ratio has statistically significant influence for families with relatively small household size on response rate.
4. The OLS model shows that dollars given not including match is influenced by matching gifts for male subjects, subjects from families with relatively low median household income at 10% level.

### 5.2 Computer Codes

Computer codes are written in Python and R. Most of the codes are written in Python, including the FWER control algorithm, and multiple testing analysis. The Probit models and OLS linear models are written in R.

The codes for the project are uploaded at here.

## References

[Bon35] Carlo E Bonferroni. “Il calcolo delle assicurazioni su gruppi di teste”. In: *Studi in onore del professore salvatore ortu carboni* (1935), pp. 13–60.

[KL07] Dean Karlan and John A List. “Does price matter in charitable giving? Evidence from a large-scale natural field experiment”. In: *American Economic Review* 97.5 (2007), pp. 1774–1793.

[KL21] Nahum Kiryati and Yuval Landau. “Dataset growth in medical image analysis research”. In: *Journal of imaging* 7.8 (2021), p. 155.

[LSX19] John A List, Azeem M Shaikh, and Yang Xu. “Multiple hypothesis testing in experimental economics”. In: *Experimental Economics* 22.4 (2019), pp. 773–793.

[RW10] Joseph P Romano and Michael Wolf. “Balanced control of generalized error rates”. In: *The Annals of Statistics* 38.1 (2010), pp. 598–633.

Table 1: Reproducing the results in [LSX19]. The results are obtained using the Python version implementation.

**Panel A: Multiple Outcomes**

| Outcome | Absolute Difference | $p$-values | $p$-values corrected with 1 | $p$-values corrected with Bonferroni | $p$-values corrected with Holm's |
|---|---:|---:|---:|---:|---:|
| Response Rate | 0.0042 | 0.0000*** | 0.0000*** | 0.0000*** | 0.0000*** |
| Dollars Given Not Including Match | 0.1536 | 0.0500* | 0.0970* | 0.2000 | 0.1000 |
| Dollars Given Including Match | 2.0876 | 0.0000*** | 0.0000*** | 0.0000*** | 0.0000*** |
| Amount Change | 6.3306 | 0.7200 | 0.7200 | 1.0000 | 0.7200 |

**Panel B: Multiple Subgroups**

| Subgroups | Absolute Difference | $p$-values | $p$-values corrected with 1 | $p$-values corrected with Bonferroni | $p$-values corrected with Holm's |
|---|---:|---:|---:|---:|---:|
| Red County in a Red State | 0.0095 | 0.0003*** | 0.0013*** | 0.0013*** | 0.0013*** |
| Blue County in a Red State | 0.0070 | 0.0503* | 0.1430 | 0.2013 | 0.1510 |
| Red County in a Blue State | 0.0000 | 0.9920 | 0.9920 | 1.0000 | 0.9920 |
| Blue County in a Blue State | 0.0015 | 0.4880 | 0.7360 | 1.0000 | 0.9760 |

**Panel C: Multiple Treatments**

| Treatment/Control Groups | Absolute Difference | $p$-values | $p$-values corrected with 1 | $p$-values corrected with Bonferroni | $p$-values corrected with Holm's |
|---|---:|---:|---:|---:|---:|
| Control vs 1:1 | 0.1234 | 0.2627 | 0.5817 | 1.0000 | 1.000000 |
| Control vs 2:1 | 0.2129 | 0.0477** | 0.1930 | 0.2860 | 0.286000 |
| Control vs 3:1 | 0.1245 | 0.2060 | 0.5540 | 1.0000 | 1.000000 |
| 1:1 vs 2:1 | 0.0895 | 0.4627 | 0.7463 | 1.0000 | 1.000000 |
| 1:1 vs 3:1 | 0.0011 | 0.9920 | 0.9920 | 1.0000 | 0.992000 |
| 2:1 vs 3:1 | 0.0883 | 0.4633 | 0.6960 | 1.0000 | 0.926667 |

[1] *, ** and *** indicates that the corresponding p-values less than 10%, 5% and 1%, respectively.

[2] The results are slightly different from the results shown in [LSX19] due to 1) the dataset mismatch and 2) the quantile calculation procedure, yet the conclusions are the same.

Table 2: Reproducing the results in [LSX19]. The results are obtained using the Python version implementation. There are 48 null hypotheses to be tested.

| Outcome | Subgroups | Treat./Cont. Groups | Absolute Difference | $p$-values | $p$-values, eqn. 1 | $p$-values, Bonf. | $p$-values, Holm's |
|---|---|---|---|---|---|---|---|
| Response Rate | RCRS | Control vs 1:1 | 0.0079 | 0.0216** | 0.4703 | 1.000 | 0.7583 |
| | | Control vs 2:1 | 0.0099 | 0.0017*** | 0.0583* | 0.080* | 0.0717* |
| | | Control vs 3:1 | 0.01067 | 0.0017*** | 0.0593* | 0.080* | 0.0733* |
| | BCRS | Control vs 1:1 | 0.0023 | 0.5973 | 1.0000 | 1.000 | 1.0000 |
| | | Control vs 2:1 | 0.0080 | 0.0987 | 0.9010 | 1.000 | 1.0000 |
| | | Control vs 3:1 | 0.0108 | 0.0247** | 0.5053 | 1.000 | 0.8387 |
| | RCBS | Control vs 1:1 | 0.0006 | 0.8667 | 1.0000 | 1.000 | 1.0000 |
| | | Control vs 2:1 | 0.0026 | 0.5033 | 1.0000 | 1.000 | 1.0000 |
| | | Control vs 3:1 | 0.0032 | 0.3740 | 1.0000 | 1.000 | 1.0000 |
| | BCBS | Control vs 1:1 | 0.0000 | 0.9053 | 0.9987 | 1.000 | 1.0000 |
| | | Control vs 2:1 | 0.0008 | 0.7877 | 1.0000 | 1.000 | 1.0000 |
| | | Control vs 3:1 | 0.0034 | 0.2313 | 0.9953 | 1.000 | 1.0000 |
| Dollars Given Not Including Match | RCRS | Control vs 1:1 | 0.4260 | 0.09033 | 0.9017 | 1.000 | 1.0000 |
| | | Control vs 2:1 | 0.4097 | 0.0557* | 0.7883 | 1.000 | 1.0000 |
| | | Control vs 3:1 | 0.3214 | 0.0710 | 0.8493 | 1.000 | 1.0000 |
| | BCRS | Control vs 1:1 | 0.0374 | 0.8950 | 1.0000 | 1.000 | 1.0000 |
| | | Control vs 2:1 | 0.4325 | 0.1853 | 0.9853 | 1.000 | 1.0000 |
| | | Control vs 3:1 | 0.5728 | 0.0933* | 0.8987 | 1.000 | 1.0000 |
| | RCBS | Control vs 1:1 | 0.0074 | 0.9747 | 0.9747 | 1.000 | 0.9747 |
| | | Control vs 2:1 | 0.0380 | 0.8650 | 1.0000 | 1.000 | 1.0000 |
| | | Control vs 3:1 | 0.2173 | 0.2847 | 0.9993 | 1.000 | 1.0000 |
| | BCBS | Control vs 1:1 | 0.0255 | 0.8680 | 1.0000 | 1.000 | 1.0000 |
| | | Control vs 2:1 | 0.0796 | 0.6430** | 1.0000 | 1.000 | 1.0000 |
| | | Control vs 3:1 | 0.0236 | 0.8880 | 1.0000 | 1.000 | 1.0000 |
| Dollars Given Including Match | RCRS | Control vs 1:1 | 1.5042 | 0.0080*** | 0.2230 | 0.3840 | 0.3040 |
| | | Control vs 2:1 | 2.5335 | 0.0010*** | 0.0363** | 0.048** | 0.0450** |
| | | Control vs 3:1 | 3.2419 | 0.0000*** | 0.0000*** | 0.000*** | 0.000*** |
| | BCRS | Control vs 1:1 | 0.8370 | 0.0603* | 0.8053 | 1.000 | 1.0000 |
| | | Control vs 2:1 | 2.8217 | 0.0080*** | 0.2250 | 0.384 | 0.3120 |
| | | Control vs 3:1 | 4.5776 | 0.0023*** | 0.0737* | 0.112 | 0.0933* |
| | RCBS | Control vs 1:1 | 0.9967 | 0.0133** | 0.3290 | 0.640 | 0.4800 |
| | | Control vs 2:1 | 2.1371 | 0.0023*** | 0.0747* | 0.112 | 0.0957* |
| | | Control vs 3:1 | 2.1658 | 0.0020*** | 0.0680* | 0.096* | 0.0840* |
| | BCBS | Control vs 1:1 | 0.7767 | 0.0087* | 0.2370 | 0.416 | 0.320667 |
| | | Control vs 2:1 | 1.8941 | 0.0007*** | 0.0247** | 0.032** | 0.0307** |
| | | Control vs 3:1 | 2.5773 | 0.0000*** | 0.0000*** | 0.000*** | 0.0000*** |
| Amount change | RCRS | Control vs 1:1 | 1.8252 | 0.1310 | 0.9497 | 1.000 | 1.0000 |
| | | Control vs 2:1 | 0.5491 | 0.6443 | 1.0000 | 1.000 | 1.0000 |
| | | Control vs 3:1 | 0.0681 | 0.9593 | 0.9983 | 1.000 | 1.0000 |
| | BCRS | Control vs 1:1 | 92.3221 | 0.4410 | 1.0000 | 1.000 | 1.0000 |
| | | Control vs 2:1 | 93.7227 | 0.4410 | 1.0000 | 1.000 | 1.0000 |
| | | Control vs 3:1 | 94.2640 | 0.4410 | 1.0000 | 1.000 | 1.0000 |
| | RCBS | Control vs 1:1 | 0.9294 | 0.4617 | 1.0000 | 1.000 | 1.0000 |
| | | Control vs 2:1 | 0.2938 | 0.8277 | 1.0000 | 1.000 | 1.0000 |
| | | Control vs 3:1 | 0.5147 | 0.6577 | 1.0000 | 1.000 | 1.0000 |
| | BCBS | Control vs 1:1 | 52.1438 | 0.4530 | 1.0000 | 1.000 | 1.0000 |
| | | Control vs 2:1 | 0.4714 | 0.6607 | 1.0000 | 1.000 | 1.0000 |
| | | Control vs 3:1 | 1.16140 | 0.2457 | 0.9953 | 1.000 | 1.0000 |

[1] *, ** and *** indicates that the corresponding p-values less than 10%, 5% and 1%, respectively.
[2] The results are slightly different from the results shown in [LSX19] due to 1) the dataset mismatch and 2) the quantile calculation procedure, yet the conclusions are the same.
[3] Code for subgroups: RC - Red County, RS - Red State, BC - Blue County, BS - Blue State.

Table 3: New insights, regression results. Probit, dependent variable = Responded (binary). The values shows the marginal effect of each variable.

| | Female | | Male | | High MHI[1] | | Low MHI[1] | | Large AHS[1] | | Small AHS[1] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| Treatment | 0.0029 | -0.0026 | 0.0051*** | 0.0047 | 0.0017 | -0.0020 | 0.0071*** | 0.0066* | 0.0034 | 0.0039 | 0.0054** | 0.0012 |
| | (0.0026) | (0.0051) | (0.0015) | (0.0026) | (0.0019) | (0.0037) | (0.0019) | (0.0032) | (0.0034) | (0.0019) | (0.0018) | (0.0034) |
| Treatment * Medium Example Amount | | 0.0009 | | 0.0006 | | -0.0007 | | 0.0020 | | -0.0001 | | 0.0016 |
| | | (0.0039) | | (0.0022) | | (0.0026) | | (0.0028) | | (0.0027) | | (0.0027) |
| Treatment * High Example Amount | | 0.0056 | | -0.0001 | | 0.0026 | | 0.0027 | | -0.0004 | | 0.0027 |
| | | (0.0042) | | (0.0022) | | (0.0026) | | (0.0028) | | (0.0027) | | (0.0027) |
| Treatment * 2:1 ratio | | 0.0034 | | 0.0014 | | 0.0020 | | 0.0024 | | -0.0019 | | 0.0067* |
| | | (0.0040) | | (0.0023) | | (0.0028) | | (0.0028) | | (0.0026) | | (0.0030) |
| Treatment * 3:1 ratio | | 0.0035 | | 0.0017 | | 0.0021 | | 0.0031 | | 0.0017 | | 0.0037 |
| | | (0.0040) | | (0.0023) | | (0.0028) | | (0.0028) | | (0.0028) | | (0.0029) |
| Treatment * $25,000 threshold | | 0.0023 | | -0.0021 | | 0.0018 | | -0.0026 | | -0.0011 | | -0.0003 |
| | | (0.0045) | | (0.0024) | | (0.0034) | | (0.0027) | | (0.0030) | | (0.0029) |
| Treatment * $50,000 threshold | | 0.0006 | | 0.0000 | | 0.0041 | | -0.0036 | | 0.0003 | | -0.0009 |
| | | (0.0045) | | (0.0025) | | (0.0035) | | (0.0027) | | (0.0031) | | (0.0029) |
| Treatment * $100,000 threshold | | 0.0017 | | -0.0011 | | 0.0055 | | -0.0048 | | -0.0003 | | -0.0002 |
| | | (0.0045) | | (0.0024) | | (0.0036) | | (0.0026) | | (0.0031) | | (0.0029) |

[1] MHI for median household income. AHS for average household size.

[2] Standard errors in parentheses. *, ** and *** indicates that the corresponding p-values less than 10%, 5% and 1%, respectively.

Table 4: New insights, regression results. OLS Linear Regression, dependent variable = Dollars Given Not Including Match.

| | Female | | Male | | High MHI[1] | | Low MHI[1] | | Large AHS[1] | | Small AHS[1] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| Treatment | 0.0517 | -0.1344 | 0.2034* | 0.2255 | 0.1092 | -0.1972 | 0.2242* | 0.4534* | 0.1147 | 0.2239 | 0.2167 | 0.0288 |
| | (0.1396) | (0.2564) | (0.1027) | (0.1874) | (0.1321) | (0.2416) | (0.1057) | (0.1920) | (0.1164) | (0.2117) | (0.1228) | (0.2244) |
| Treat. | | -0.1239 | | 0.1615 | | 0.0930 | | 0.0414 | | -0.0624 | | 0.2063 |
| * Medium | | (0.1996) | | (0.1448) | | (0.1871) | | (0.1489) | | (0.1638) | | (0.1742) |
| Treat. | | 0.2234 | | 0.0097 | | 0.0352 | | 0.0525 | | -0.0959 | | 0.1874 |
| * High | | (0.1988) | | (0.1450) | | (0.1870) | | (0.1488) | | (0.1632) | | (0.1747) |
| Treat. * 2:1 | | 0.0949 | | 0.1108 | | 0.2314 | | 0.0088 | | 0.0299 | | 0.2075 |
| | | (0.1980) | | (0.1453) | | (0.1876) | | (0.1486) | | (0.1632) | | (0.1750) |
| Treat. * 3:1 | | 0.3041 | | -0.0925 | | 0.0590 | | 0.0241 | | 0.0978 | | -0.0205 |
| | | (0.1995) | | (0.1448) | | (0.1864) | | (0.1492) | | (0.1635) | | (0.1742) |
| Treat. * $25k | | 0.2068 | | -0.0171 | | 0.2720 | | -0.1860 | | -0.0443 | | 0.1333 |
| threshold | | (0.2290) | | (0.1676) | | (0.2157) | | (0.1720) | | (0.1890) | | (0.2011) |
| Treat. * $50k | | -0.0388 | | -0.1700 | | 0.2032 | | -<br>0.4683** | | -0.2068 | | -0.0595 |
| threshold | | (0.2307) | | (0.1672) | | (0.2162) | | (0.1717) | | (0.1877) | | (0.2025) |
| Treat. * $100k | | -0.0834 | | -0.1517 | | 0.1948 | | -0.4322* | | -0.1426 | | -0.0957 |
| threshold | | (0.2305) | | (0.1672) | | (0.2158) | | (0.1718) | | (0.1888) | | (0.2011) |
| Constant | 0.8075*** | 0.8075*** | 0.8132*** | 0.8132*** | 0.9506*** | 0.9506*** | 0.6733*** | 0.6733*** | 0.8709*** | 0.8709*** | 0.7542*** | 0.7542*** |
| | (0.1135) | (0.1135) | (0.084) | (0.084) | (0.1078) | (0.1078) | (0.0864) | (0.0864) | (0.0953) | (0.0953) | (0.1001) | (0.1001) |
| R-squared | 0.0000 | 0.0005 | 0.0001 | 0.0002 | 0.0000 | 0.0001 | 0.0002 | 0.0006 | 0.0000 | 0.0001 | 0.0001 | 0.0003 |
| Observation | 13591 | 13591 | 35367 | 35367 | 24108 | 24108 | 24094 | 24094 | 24230 | 24230 | 23991 | 23991 |

[1] MHI for median household income. AHS for average household size.

[2] Standard errors in parentheses. *, ** and *** indicates that the corresponding p-values less than 10%, 5% and 1%, respectively.

Table 5: New insights discovered by multiple hypotheses testing. The listed values are $p$-values of the tests.

| Outcomes | Subgroups | Absolute Difference | $p$-values | $p$-values corrected with 1 | $p$-values corrected with Bonferroni | $p$-values corrected with Holm's |
|---|---|---|---|---|---|---|
| Response rate | Female | 0.0029 | 0.2617 | 0.6647 | 1.0000 | 1.0000 |
| | Male | 0.0051 | 0.0010*** | 0.0103** | 0.0120** | 0.0110** |
| | High MHI | 0.0017 | 0.3427 | 0.6470 | 1.0000 | 1.0000 |
| | Low MHI | 0.0064 | 0.0007*** | 0.0077*** | 0.0080*** | 0.0080*** |
| | Large AHS | 0.0035 | 0.0703* | 0.3083 | 0.8440 | 0.4923 |
| | Small AHS | 0.0048 | 0.0057*** | 0.0443** | 0.0680* | 0.0567* |
| Dollars Given Not Including Match | Female | 0.0517 | 0.7210 | 0.7210 | 1.0000 | 0.7210 |
| | Male | 0.2035 | 0.0510* | 0.2527 | 0.6120 | 0.4080 |
| | High MHI | 0.1092 | 0.4123 | 0.6417 | 1.0000 | 0.8247 |
| | Low MHI | 0.1957 | 0.0417** | 0.2357 | 0.5000 | 0.3750 |
| | Large AHS | 0.1148 | 0.3273 | 0.7103 | 1.0000 | 1.0000 |
| | Small AHS | 0.1885 | 0.0963* | 0.3757 | 1.0000 | 0.5780 |

*, **, and *** indicate that the corresponding p-value less than 10%. 5%, and 1% respectively.